EASTERN REGION TECHNICAL ATTACHMENT
NO. 91-7B
AUGUST, 1991

A NEW VERIFICATION SCHEME

George J. Maglaras
National Weather Service Forecast Office
Albany, New York

1. INTRODUCTION

The National Weather Service (NWS) at Albany, New York, (WSFO ALB), has recently implemented a new verification procedure. This is done in addition to, and utilizes data from, the Western Region local verification program (Barker 1987), which is run on AFOS. We did this to inspire greater forecaster interest and possibly promote some friendly competition. The new scheme is based on the premise that one of the most desirable overall verification measures is to determine how frequently forecasters deviate substantially from MOS guidance, and how effective they are when they do so. In order to take into account both how frequently and how effectively forecasters deviate from MOS, a new score has been formulated. This score is called Frequently and Effectively Departs Significantly (FEDS).

For each forecaster, the FEDS score is calculated for an entire cool (October-March) or warm (April-September) season for both the maximum/minimum temperature and probability of precipitation (PoP) forecasts, separately; and a combined temperature/PoP score is also determined. In addition, a long term average FEDS score is calculated for all previous seasons combined, similar to a lifetime batting average for a baseball player.

2. BACKGROUND AND DEFINITION

Since the implementation of the AFOS-era Verification program (Dagostaro 1985), several verification scores have been provided to individual forecasters and forecast offices in order to help assess forecaster performance. For temperature, the score used to directly measure forecaster error is the mean absolute error (MAE), and for PoP it is the NWS Brier score (half the score originally proposed by Brier 1950). The MAE and Brier scores are used as the basis for determining the overall percent improvement over MOS. The percent improvement over MOS is one of the measures used for the determination of overall forecaster effectiveness.

For temperature, several other scores are calculated. A threat score is determined for those cases when the observed, or MOS, or the local temperature forecasts changed by 10°F, or more, from one day to the next. Another score is the MAE for those cases when either MOS or the local forecast temperatures had an error of 6°F, or more. Finally, the significant change score is the MAE for those cases where the local forecast deviated from MOS by 3°F, or more. For each of these, the percent improvement over MOS is also calculated to indicate forecaster effectiveness. Analogously, for the PoP forecasts, the significant change score is the percent improvement over the MOS PoPs for those cases when a forecaster has deviated from MOS PoP by 20% or more.

It is my opinion that the improvement over MOS for the significant change score is the most important of the calculations, and probably more important than the overall improvement over MOS as well. Some of my reasons are: it is not common for a forecaster to keep track of cases where the threat or error scores apply, but forecasters almost always know when they have deviated substantially from MOS; for forecasters who take pride in their verification scores, it is most important that they improve over MOS when they have made substantial departures; and this new score can help motivate forecasters. The magnitude of the improvement over MOS for the significant change score is frequently in the 20% to 40% range and usually two to four times greater than the overall improvement over MOS. This shows forecasters that apply sound meteorological reasoning that they can make a big difference in the forecast for many important weather events. In contrast, for the overall improvement over MOS, a forecaster may be tempted to say: "Why bother trying to beat MOS if I only improve by 5% or 10%?"

The Western Region local verification program provides output that shows both how frequently (in percent) forecasters deviate significantly from MOS, and also their improvement over MOS when they do so. The FEDS score is calculated by multiplying the frequency of significant changes (F) by the improvement over MOS (I), and then dividing by ten. To this total the overall percent improvement over MOS (OI) is then added. Hence:

$$\text{FEDS} = ((\text{F} \times \text{I}) / 10) + \text{OI}.$$

Forecasters who deviate both frequently and effectively from MOS, and also have a meaningful overall improvement over MOS, will have the best FEDS scores. Of course, forecasters who deviate effectively from MOS, but only do so for sure things (low frequency), will not have a high FEDS score. Analogously, forecasters who depart significantly from MOS many times, but are not skillful (low improvement over MOS), will also have low FEDS scores. In addition, forecasters who have little or no overall improvement over MOS also are likely to have lower FEDS scores.

3. THE NEW SCORE

Tables 1 and 2 show the individual FEDS, and overall percent improvement over MOS scores, for both temperature and PoP for the 1990-91 cool and 1990 warm seasons, respectively. Also provided is the combined temperature/PoP FEDS score (temperature FEDS + PoP FEDS). The number of forecasts given in these Tables is the number of Coded Cities Forecast (CCF) products issued by a forecaster multiplied by the number of stations in the CCF. For example, in Table 1, forecaster A issued the CCF product 11 times during the cool season. Since the ALBCCFALB product includes forecasts for two cities, forecaster A had 22 forecasts overall. The overall percent improvement over MOS is included in these tables so that forecasters can determine how much this value contributed to the FEDS score. It also helps to point out that some forecasters with relatively small average improvements over MOS, may still have substantial FEDS scores.

Table 3 shows the lifetime average FEDS scores for temperature, PoP, and temperature/PoP combined. Also included in Table 3 is the number of seasons for which the FEDS score was calculated and the number of forecasts each forecaster made over his/her lifetime.

4. DISCUSSION

Although I would not recommend that it be used exclusively, the new verification scheme implemented at WSFO ALB has several noteworthy features. First, the emphasis is on a verification measure that is meaningful to most forecasters, namely, beating MOS when they have departed from it in a substantial manner. Second, forecasters need only look at a single number, the seasonal FEDS score for temperature/PoP combined, to get an important indication of their overall success. Third, I think this approach will encourage forecasters to depart substantially from guidance by learning to recognize weather regimes where either the MOS guidance, or the numerical models that produce MOS, do not perform well. Once forecasters have identified the types of weather regimes (either long term anomalous weather patterns or short term single event cases) that consistently produce large MOS errors, they can confidently depart substantially from MOS for those cases with a high probability of success. Fourth, the lifetime FEDS scores, which are easy to calculate and carry over from season to season, can be used by new forecasters to monitor their improvement as they gain experience and maturity. Finally, and most important, I believe the use of FEDS will result in greater forecaster interest in the verification system, and may promote some friendly competition, similar to the manner in which sports fans have fun discussing sports statistics. At the end of your career, when you retire from a station, the lifetime FEDS score can be used as a basis for retiring your forecaster number. Or, if you go to another office someday, your agent can use this score to get you a couple of extra million bucks for your contract.

5. ACKNOWLEDGMENTS

I greatly appreciate the thoughtful reviews and suggestions for improvement that Dave Unger, Techniques Development Laboratory, NWS Office of Systems Development, provided for previous versions of this manuscript.

REFERENCES

Barker, T.W., 1987: AEV local verification for aviation, precipitation, and temperature programs. NOAA Western Region Computer Programs and Problems NWS WRCP-42, NOAA/NWS, Salt Lake City, UT, 33 pp.

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1-3.

Dagostaro, V.J., 1985: The national AFOS era verification processing system. TDL Office Note 85-9, NOAA/NWS, Silver Spring, MD, 47 pp.

FCSTR	# FCSTS	Temperature FEDS	Temperature % IMPROV	PoP FEDS	PoP % IMPROV	Temp/PoP FEDS
C	70	93.1	12.5	70.7	11.4	163.8
A	22	178.9	22.1	-30.7	-1.2	148.2
B	84	111.1	13.2	18.5	4.6	129.6
K	44	31.5	5.7	74.6	11.7	106.1
F	52	56.3	8.9	37.3	8.2	93.6
D	70	79.6	9.9	8.6	4.0	88.2
G	84	22.9	7.5	39.4	10.9	62.3
J	56	76.8	6.6	-67.3	-6.5	9.5
I	38	110.8	13.4	-123.9	-18.1	-13.1
H	56	-13.6	-1.4	-41.2	-7.3	-54.8
E	54	-37.4	-3.2	-71.5	-11.0	-108.9
Station	654	62.1	8.2	6.3	2.4	68.4

Table 1. Individual FEDS and overall percent improvement over MOS scores for temperature and PoP for the 1990-91 cool season, and the combined temperature/PoP FEDS score.

FCSTR	# FCSTS	Temperature FEDS	Temperature % IMPROV	PoP FEDS	PoP % IMPROV	Temp/PoP FEDS
I	70	84.7	10.7	53.6	7.3	138.3
B	79	69.4	8.9	53.5	10.4	122.9
A	18	31.8	4.2	57.0	13.3	88.8
G	54	36.6	7.5	19.2	7.3	55.8
J	84	41.7	7.9	14.1	2.8	55.8
C	70	50.0	7.4	-1.3	2.1	48.7
E	82	25.5	0.9	15.6	0.6	41.1
D	48	12.0	-0.4	17.3	6.2	29.3
H	64	30.5	3.6	-8.8	2.1	21.7
F	54	12.2	3.6	-3.6	-1.0	8.6
K	38	8.0	2.8	-37.4	-5.7	-29.4
Station	726	38.3	5.4	20.8	4.0	59.1

Table 2. Same as Table 1, except for the 1990 warm season.

FCSTR	# FCSTS	Temperature FEDS	PoP FEDS	Temp/PoP FEDS	# of SEASONS
B	163	90.9	34.5	126.4	2
A	40	112.7	8.8	121.5	2
C	140	71.6	34.7	106.3	2
I	108	93.9	-8.9	85.0	2
D	118	52.1	12.1	64.2	2
G	138	28.3	31.5	59.8	2
F	106	33.8	16.5	50.3	2
K	82	20.6	22.7	43.3	2
J	140	55.7	-18.5	37.2	2
H	120	9.9	-23.9	-14.0	2
E	136	0.5	-19.0	-18.5	2
Station	1380	49.6	13.9	63.5	2

Table 3. Same as Table 1, except only for the lifetime FEDS averages.